

SENT BY: ;

4084679681;

JUL-8-04 3:03PM;

PAGE 2/2

PATENT
AM9-99-074IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:)
ANITA WAI-LING HUANG, et al.) Group Art Unit: 2178
Serial No.: 09/513,058)
Filed: February 24, 2000) Examiner: A. Baschoar
For: SYSTEM AND METHOD FOR)
CLASSIFYING ELECTRONICALLY)
POSTED DOCUMENTS)

37 C.F.R. § 1.131 DECLARATION

We, the undersigned, are the Applicants for the above-identified patent application and hereby declare the following:

- 1) The pending claims of our above-identified patent application were rejected under 35 U.S.C. § 103(a) based on U.S. Patent No. 5,913,208 to Brown et al., which is entitled "Identifying Duplicate Documents from Search Results Without Comparing Document Content" and issued on June 15, 1999 ("Brown").
- 2) The invention claimed in the above-identified patent application was reduced to writing in the United States prior to the June 15, 1999 issue date of the Brown reference. Attached hereto is the relevant portion of an Invention Disclosure on which the above-identified patent application was based. This Invention Disclosure was prepared prior to June 15, 1999.

We, the undersigned, hereby declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Name: Anita Wai-Ling Huang*

Signature: _____

Date: _____


Name: Neelakantan Sundaresan

Signature: 

Date: 7/7/04

*Unavailable for signature under MPEP § 715.04: no longer employed by IBM and unreachable at last known mailing addresses, email addresses, and phone numbers.

A System and Method Based on Metadata for Eliminating Duplicates in Result Sets in Internet Search Engines

	Disclosure ARC8-1999-0080	
	Created By: Neel Sundaresan	Created On:
	Last Modified By: Neel Sundaresan	Last Modified:
	*** IBM Confidential ***	

Required fields are marked with the asterisk (*) and must be filled in to complete the form.

Summary

Status	Under Evaluation
Processing Location	ARC
Functional Area	OPB - Computer Science - (A.K. Chandra)
IDT Team	Khanh Tran/Almaden/IBM; Cheryl Ruby/Almaden/IBM
Submitted Date	
Owning Division	RES
PVT Score	45

Inventors with Lotus Notes ID's

Inventors: Anita Huang/Almaden/IBM, Neel Sundaresan/Almaden/IBM

Inventor Name	Inventor Serial	Div/Dept	Manager Serial	Manager Name
> denotes primary contact				
Huang, Anita W.	880805	22/K575	080852	Ford, Daniel A.
> Sundaresan, Neelakantan (Neel)	883678	22/K575	080852	Ford, Daniel A.

Inventors without Lotus Notes ID's

IDT Selection

IDT Team:	Attorney/Patent Professional
Khanh Tran/Almaden/IBM	
Cheryl Ruby/Almaden/IBM	

Response Due to IP&L: 6

Main Idea

***Title of disclosure (in English):**
A System and Method Based on Metadata for Eliminating Duplicates in Result Sets in Internet Search Engines.

*Idea of disclosure

1. Describe your invention, stating the problem solved (if appropriate), and indicating the advantages of using the invention.
Search engines of today have the problem that they cannot differentiate between the same data at different sites. Both the pieces of data show up on search results only because they are from different web sites. It is an annoyance to end user if this data qualifies for a particular search request and all the URLs where this data is duplicated shows up. This is particularly true when popular web documents are mirrored at several web sites.

We propose a solution based on metadata to eliminate such duplicates. Our web crawler crawls the web and builds metadata in XML-encoded RDF form. The metadata for a web source mainly

A System and Method Based on Metadata for Eliminating Duplicates in Result Sets in Internet Search Engines

contains three pieces of information -1. information about the crawler, crawl date etc., 2. information about the data source from where the data was obtained., 3. information about the data itself. All of these are kept in XML/RDF form. The advantage of keeping it in such a form is that we keep structural information around and hence our metadata is of high quality. The metadata is kept in a metadata repository before it is fed to an index engine which serves a search engine. Our invention will work in the metadata repository to identify duplicates. For each piece of metadata it will try and compare other pieces of metadata with this one. If the other piece of metadata qualifies as equal to this one, it will eliminate it from the repository. This way duplicates are eliminated.

In order to avoid exhaustive (quadratic) search of the metadata repository, the system will automatically eliminate those that are obviously different from comparison. For instance the metadata for a Java program will contain a tag saying that it is a Java program. Obviously this piece of metadata will not be compared to one that corresponds to an XML data or a C++ program.

2. How does the invention solve the problem or achieve an advantage, (a description of "the invention", including figures inline as appropriate)?

Block Diagrams:

A System and Method Based on Metadata for Eliminating Duplicates in Result Sets in Internet Search Engines

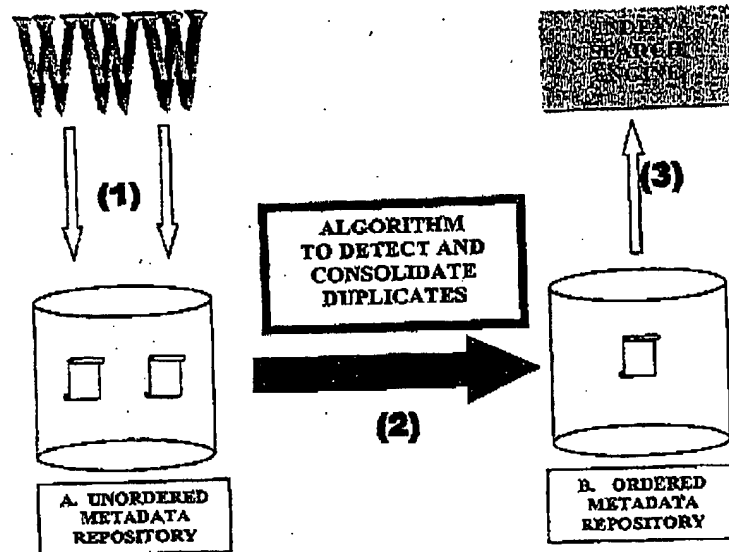


Figure 1

1. The crawlers summarize data from the World Wide Web (WWW), in RDF format, storing the summaries in a metadata repository (A). The repository contains distinct summaries for each URL.
2. Our algorithm analyzes the metadata in (A), taking advantage of structural information, for duplicated data. During this process, it consolidates groups of duplicated instances (with distinct URLs) into single records. Each consolidated record also retains a list of the duplicating URLs. The result is an ordered metadata repository (B).
3. The search engine indexes and queries the new metadata repository (B). As a result, it returns a single record for each set of duplicated instances rather than separate records for each.

A System and Method Based on Metadata for Eliminating Duplicates in Result Sets in Internet Search Engines

Sample Metadata Summaries of Identical HTML Source Files

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/schemas/rdf-schema"
  xmlns:abc="file://www.ABC.com/abc-schema">
  <rdf:Description
    abc:gatherer="TheGatherer"
    abc:gathered-on="Tue Mar 23 17:38:40 GMT 1999"
    abc:resource="http://www.foo.com/bar.html"
    abc:summarizer="TheSummaryMaker"
    abc:source-last-modified="[not known]"
    abc:mime-type="http/html"
    abc:source-is="Good"
    abc:comments="good"/>
  <rdf:Description
    abc:html-title="Duplicate Example"
    abc:html-encoding="8859_1"
    abc:abstract="Duplicate Example!"
    rdf:HTTP="">
  <abc:ref-annotations>
  <rdf:Bag>
  <rdf:LI>
    <rdf:Description
      annotates="http://foobar.com/"
      annotation="... FOO"/>
    </rdf:LI>
  </abc:ref-annotations>
  <abc:presentation-text>
  <rdf:Bag>
  <rdf:LI> Welcome to the FOOBAR Example! </rdf:LI>
  </rdf:Bag>
  </abc:presentation-text>
  </rdf:Description>
</rdf:RDF>
</abc:presentation-text>
</rdf:Description>
</rdf:RDF>
```

```
<?xml version="1.0"?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/schemas/rdf-schema"
  xmlns:abc="file://www.ABC.com/abc-schema">
  <rdf:Description
    abc:gatherer="TheGatherer"
    abc:gathered-on="Tue Mar 23 17:38:40 GMT 1999"
    abc:resource="http://www.bar.com/foo.html"
    abc:summarizer="TheSummaryMaker"
    abc:source-last-modified="[not known]"
    abc:mime-type="http/html"
    abc:source-is="Good"
    abc:comments="good"/>
  <rdf:Description
    abc:html-title="Duplicate Example"
    abc:html-encoding="8859_1"
    abc:abstract="Duplicate Example!"
    rdf:HTTP="">
  <abc:ref-annotations>
  <rdf:Bag>
  <rdf:LI>
    <rdf:Description
      annotates="http://foobar.com/"
      annotation="... FOO"/>
    </rdf:LI>
  </abc:ref-annotations>
  <abc:presentation-text>
  <rdf:Bag>
  <rdf:LI> Welcome to the FOOBAR Example! </rdf:LI>
  </rdf:Bag>
  </abc:presentation-text>
  </rdf:Description>
</rdf:RDF>
</abc:presentation-text>
</rdf:Description>
</rdf:RDF>
```

Figure 2

This example shows the metadata summaries for two URLs: <http://www.foo.com/bar.html> and <http://www.bar.com/foo.html>. Each URL references a distinct file. As the metadata shows, however, the files are identical. Our algorithm uses this metadata information to detect duplications and to consolidate two or more such records into a single record. By using metadata, rather than the actual data, our algorithm achieves an advantage over slower text comparison algorithms. Moreover, it is able to retain additional useful information, like all duplicating URLs, without cluttering the search space.

Our method is better in that it does structural comparison. Instead of doing textual comparison it does structural comparison. It also has automatic method of eliminating obvious documents that should not be compared based upon some key attribute values (like file name extensions etc.)

3. If the same advantage or problem has been identified by others (Inside/outside IBM), how have those others solved it and does your solution differ and why is it better?

A System and Method Based on Metadata for Eliminating Duplicates In Result Sets In Internet Search Engines

Alta vista solves this problem by text comparison which is a significantly slower process. On our search we found that they did not have a fool proof method for eliminating duplicates. Same with Infoseek and hotbot.

4. If the invention is implemented in a product or prototype, include technical details, purpose, disclosure details to others and the date of that implementation.
Being incorporated in the Grand Central Family of search engines and portals.